Linear Regression

CS434

Supervised learning
A regression problem

• We want to learn to predict a person’s height based on his/her knee height and/or arm span.

• This is useful for patients who are bed bound and cannot stand to take an accurate measurement of their height.

• Training data:
  – The measurements that you have taken.

Supervised learning
Target function

• Let $y$ represent a person’s height, and $\mathbf{x}$ represents the measurements we would use to predict $y$

• Here $\mathbf{x}$ contains two measurements—referred to as features, knee height and arm span

\[
\mathbf{x} = \begin{bmatrix}
    x_1 \\
    x_2
\end{bmatrix}
\]

  - $x_1$: Knee height
  - $x_2$: Arm span
Target Function Representation

• Linear function (thus the name *linear regression*):
  \[ y = w_1 x_1 + w_2 x_2 + b = \mathbf{w}^T \mathbf{x} + b \]

• Let’s start with just one feature \( x_1 \), and call it \( x \) and we want to learn a function
  \[ y = wx + b \]
error = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y})^2

\Rightarrow \frac{\partial e}{\partial b} = -\frac{2}{n} \sum_{i=1}^{n} (y_i - wx_i - b) = 0

\Rightarrow \frac{1}{n} \sum_{i=1}^{n} (y_i - wx_i - b) = 0

\Rightarrow b = \frac{1}{n} \sum_{i=1}^{n} y_i - \frac{1}{n} \sum_{i=1}^{n} x_i

b = \bar{y} - \bar{x}

\Rightarrow e(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - wx_i - \bar{y} - \bar{x})^2

= \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y} - w(x_i - \bar{x}))^2

\Rightarrow \frac{\partial e}{\partial w} \sum_{i=1}^{n} [y_i - \bar{y} - w(x_i - \bar{x})] (x_i - \bar{x}) = 0

\Rightarrow \sum_{i=1}^{n} (y_i - \bar{y}) (x_i - \bar{x}) - w \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0

\Rightarrow w = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}

b = \bar{y} - w \bar{x}

\Rightarrow W^* = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}

\Rightarrow b^* = \bar{y} - w^* \bar{x}

We can remove the need for b by centering the data:

\hat{y}_i = y_i - \bar{y}

\hat{x}_i = x_i - \bar{x}

After centering \bar{X} = \bar{Y} = 0, \ b = 0.

W^* = \frac{\sum_{i=1}^{n} \hat{y}_i \hat{x}_i}{\sum_{i=1}^{n} \hat{x}_i^2}
Vector/Matrix representation:

Let \( Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \) and \( X = \begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix} \) (\( n \times 1 \))

\[ W^* = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} = \frac{X^T Y}{X^T X} = (X^T X)^{-1} X^T Y. \]

This solution holds for a more general case where \( x_i = [x_{i1}, x_{i2}, \ldots, x_{im}] \) \( \Rightarrow m \) features.

\[ Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, \quad n \times m \]

\[ W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix}, \quad n \times 1 \]

\[ x_{11} w_1 + x_{12} w_2 + \cdots + x_{1m} w_m = y_1, \]

\[ \vdots \]

\[ x_{n1} w_1 + x_{n2} w_2 + \cdots + x_{nm} w_m = y_n \]

A system of linear equations:

\[ X \cdot W = Y \]

over-constrained because \( m \ll n \).

Normal equation:

\[ X^T X W = X^T Y \]

If \( X^T X \) has a well-defined inverse then we have a solution for \( W \):

\[ W = (X^T X)^{-1} X^T Y \], linearly

If \( X \)'s columns are independent, i.e., you can't predict one feature using a linear combination of other features, then \( (X^T X)^{-1} \) exists. \( W^* \) has a unique solution.